

Coding Agents as Design Searchers

An autonomous TAGE tuning campaign for CBP-NG

Matthew Pallan · University of Maryland

CBP-NG · ISCA 2026 · Raleigh, NC

THE CONTRIBUTION IS THE METHOD, NOT THE PREDICTOR

2,471

evaluations scored

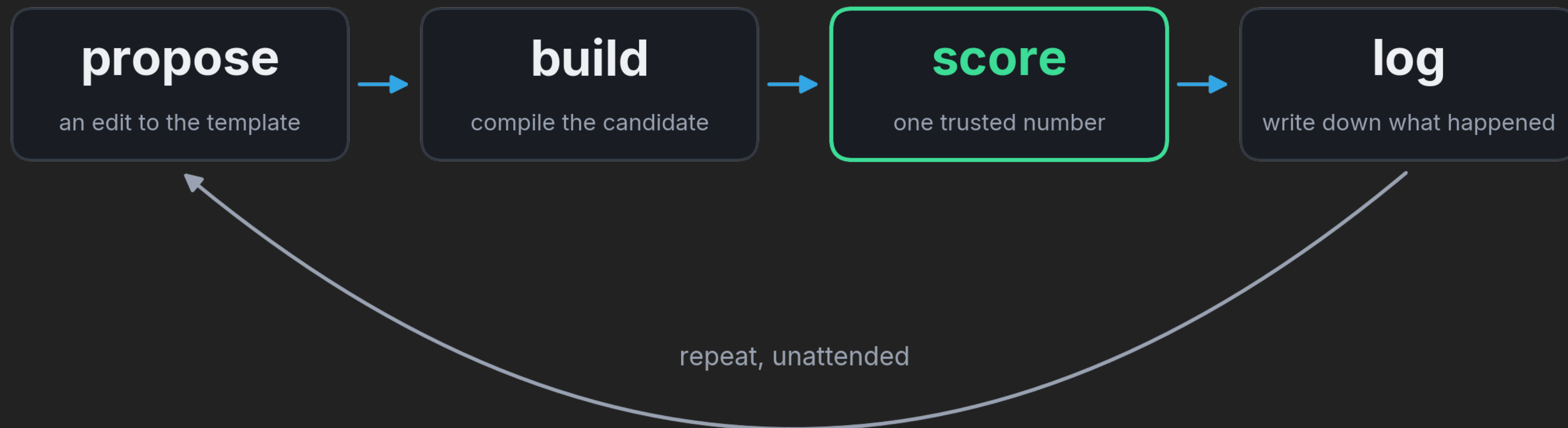
1,700+ ~3 weeks

configurations tried

running on its own

WHY BRANCH PREDICTION SUITS AUTONOMOUS AGENTS

It all hangs on one cheap, deterministic score the agent can trust.



TWO AGENTS, A THIN HUMAN ROLE

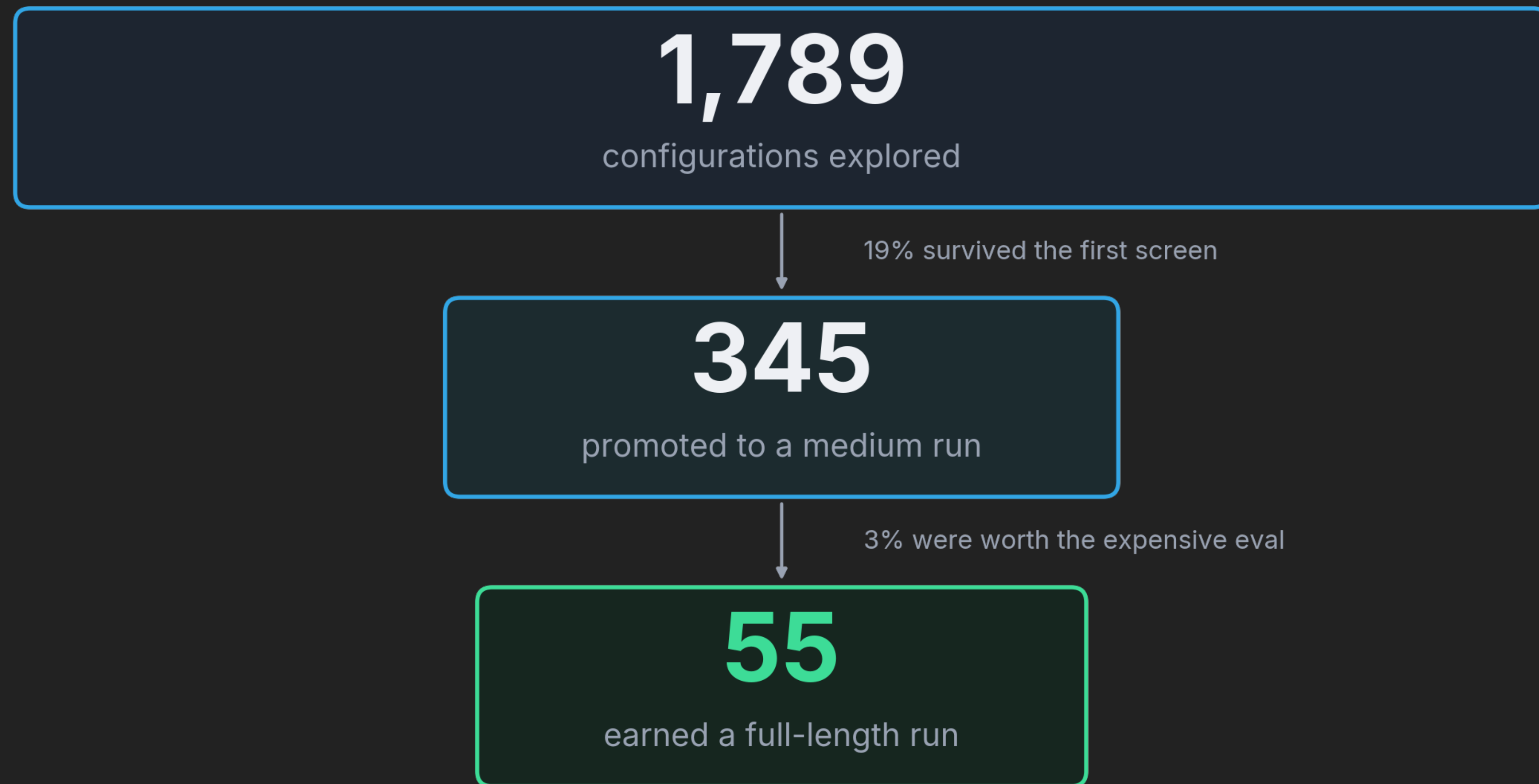
```
9b58f1e 2026-04-01 Update predictor to final VFS 0.934 config with complete journal
f1bb6e2 2026-04-01 Add .gitignore for build artifacts, binaries, traces, and sweep archives
24b9948 2026-03-24 Add tage_p1 (TAGE-in-P1 architecture) – p2_lat=1 NOT achievable
a749fc9 2026-03-23 TAGMIX=3 dead (-0.001761), METATH=10 dead (-0.001202). More structural ideas exhausted.
73abf02 2026-03-23 Batch2: WEAKTH=2/METATH=1/BIMHIST=0 all dead on 168 traces
ad39d5e 2026-03-23 WEAKTH=11 dead on 168 traces (-0.000241 VFS). 10-trace positive was misleading.
ff685dc 2026-03-23 Batch structural experiments: all dead
ec996ec 2026-03-23 TAGMIX=1 dead on 168 traces (-0.000227 VFS)
8d74371 2026-03-23 TAGMIX=1 fix: use tage_la_reg for tag hash (p1_lat safe, +0.000361 on 10 traces)
22a271f 2026-03-23 MMOM=1 COLLAPSED at full-length: -0.000195 VFS (was +0.000418 medium)
1b08d78 2026-03-23 Add MMOM=2 newly_alloc bypass (same VFS as MMOM=1 on 10 traces)
ae03cf3 2026-03-23 MMOM v2 medium 168 traces: +0.000418 VFS (0.908907 vs 0.908489)
1eb37df 2026-03-23 MMOM v2: OR momentum into metasign (was no-op with METATH=0, now +0.000747 on 10 traces)
22ec054 2026-03-23 Add 3 new mechanisms: MMOM, REFRACT, INDEXMOM (all register-based)
ccedcfc 2026-03-22 Update exploration scale: 140+ → 2300+ configurations (237 sessions)
```

THE JOURNAL: EVERY RUN, LOGGED

```
# predictor_sweeps.csv    2,532 runs    ·    400+ sweeps    ·    short → medium → full

label      vfs      notes
stock      0.79233  baseline full public sweep
tuned100    0.83703  loose larger-block tuned tage
p1lean      0.84056  first lean-P1 win
tage_sc_default  0.39319  architecture smoke vs tuned tage
tage_loop_default  0.60017  architecture smoke vs tuned tage
af2         0.90760  ALLOCFLT=2 medium  -0.000882  DEAD
sm180       0.75965  shortest-match selection  DEAD (-0.149)
mv182       0.83135  majority vote  DEAD p2_lat=3 (-0.077)
ae183       0.76110  age-based eviction  DEAD (-0.147)
base        0.93436  base for pidx29 full-length sweep
pidx29      0.93440  PATHIDX=29 full  +0.000043 (noise level)  ← final
```

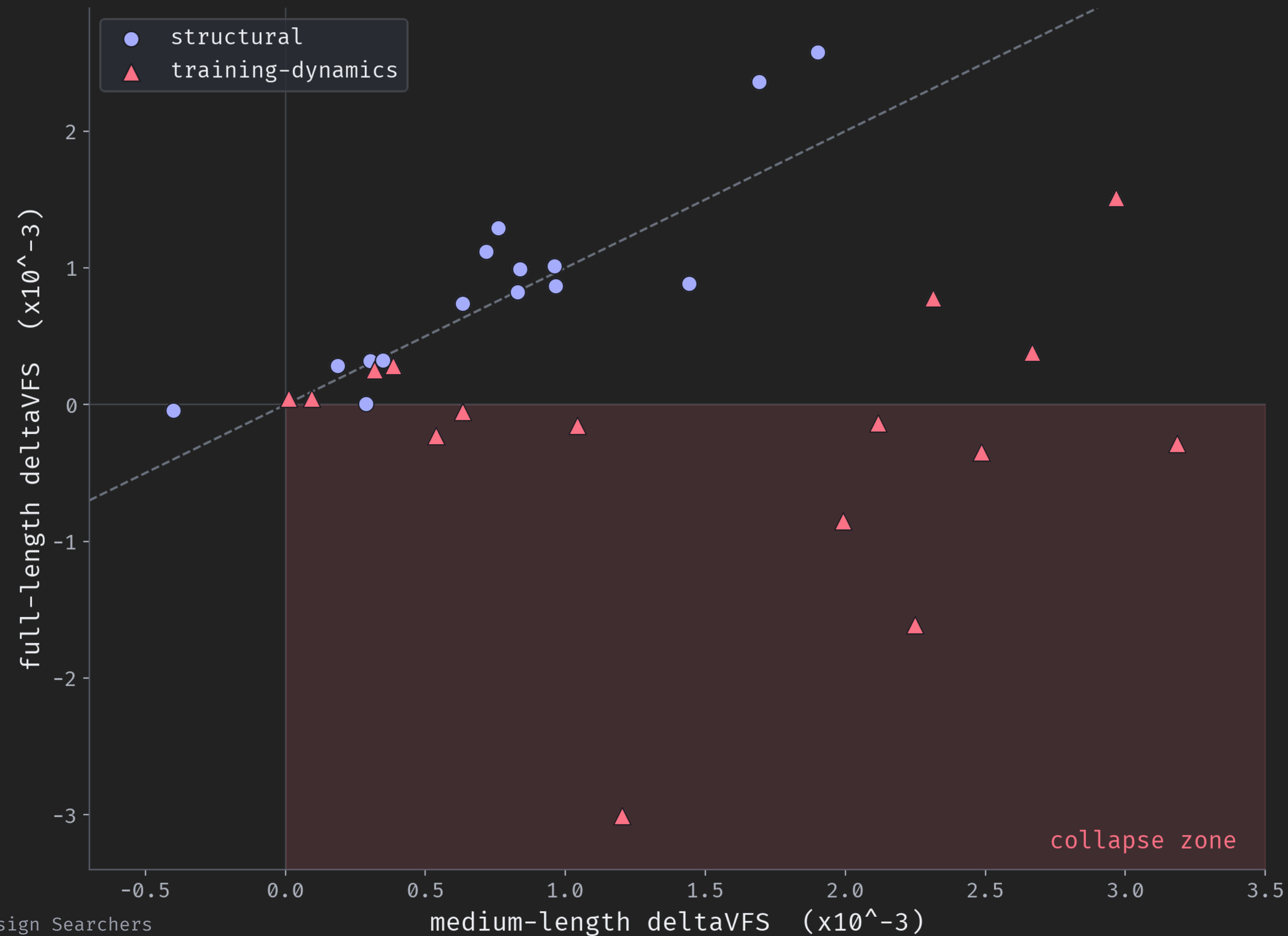
MOST IDEAS NEVER EARN A FULL RUN



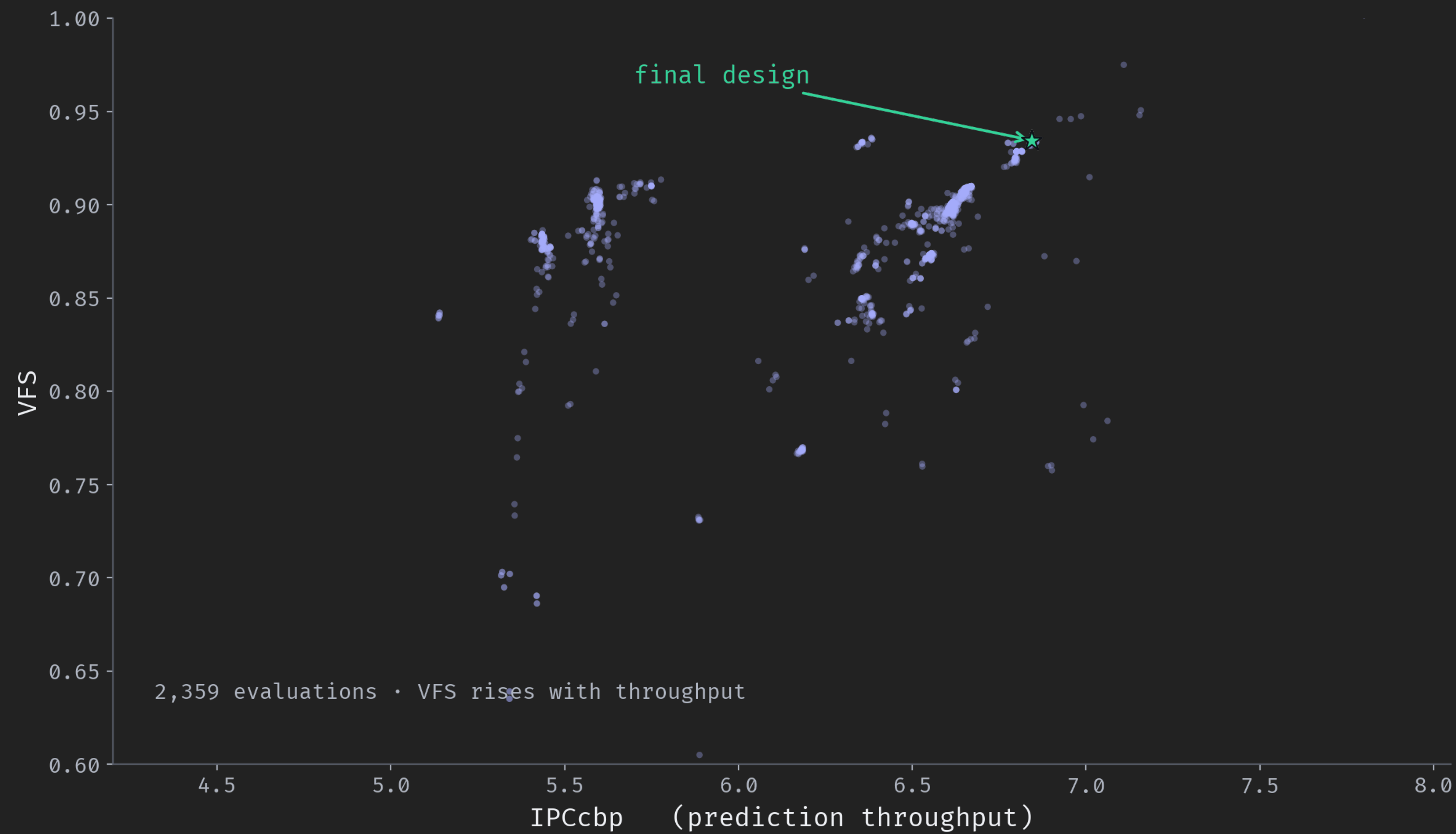
AUTONOMY IN PRACTICE



THE SEARCH FOUND ITS OWN SHORTCUT



THE RESULT: THROUGHPUT DOMINATES



WHAT THE AGENTS GOT WRONG

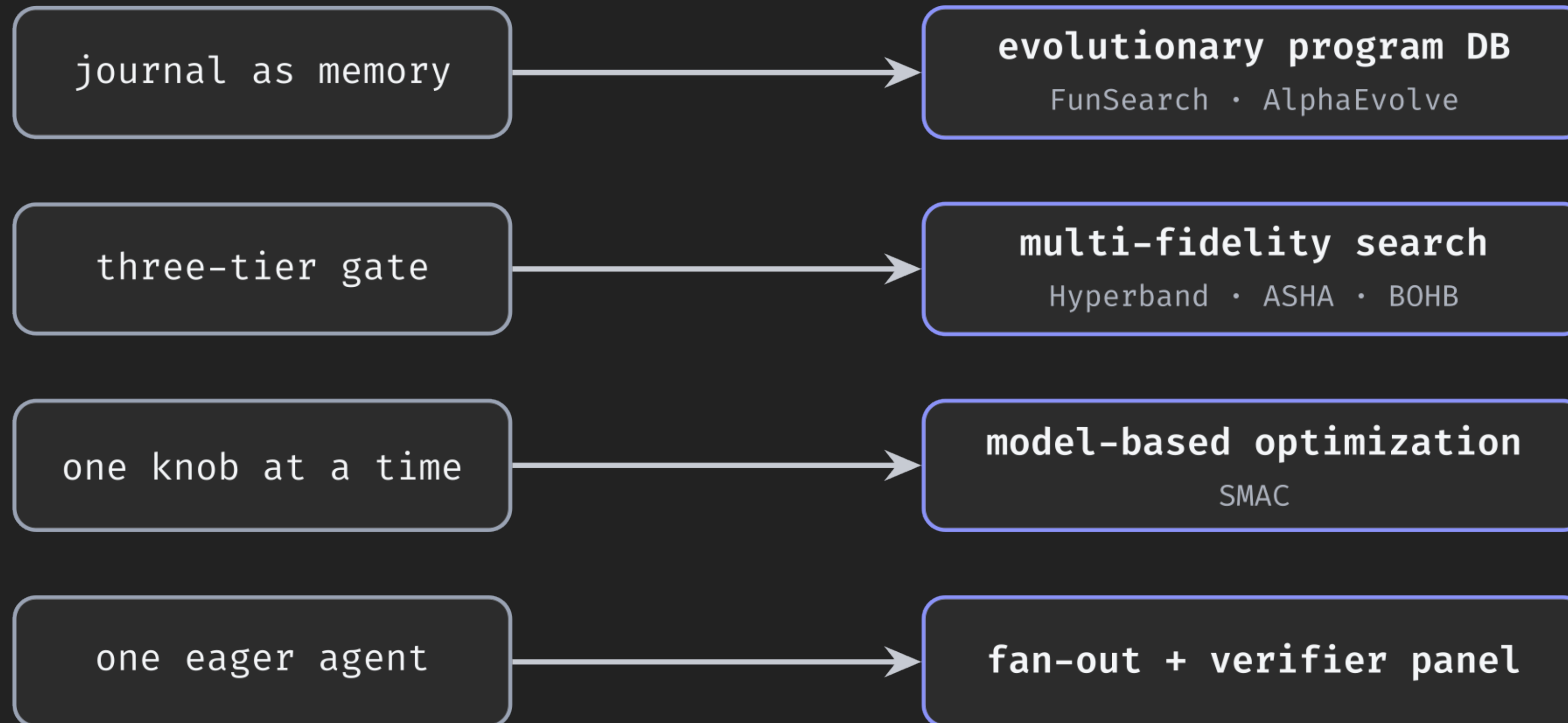
- ▶ Credulous about noise: logged sub-0.0001 'gains' until a rerun killed them
- ▶ Local search: tuned relentlessly, never invented a new architecture
- ▶ One knob at a time, even though the knobs clearly interact
- ▶ Leaned on scaffolding: an external journal and my manual restarts

Assisted autonomy. Not a scientist in a box. Not yet.

WHAT A PURPOSE-BUILT HARNESS WOULD ADD

what the agents improvised

the established method



TAKEAWAYS

- 1 Agents can run a real, multi-week hardware design search, unattended.
- 2 All they need is a cheap, trusted score they can grade themselves on.
- 3 The pattern should transfer wherever a cheap, trusted oracle exists.
- 4 Models keep improving, so the scaffolding keeps shrinking.

| Thank you

Matthew Pallan · University of Maryland

mtp@umd.edu

CBP-NG · ISCA 2026 · Raleigh, NC